The analysis and expansion of regulatory binding site data in a wide range of bacteria using semi-automatic system - RegTransBase

Pavel Novichkov, Michael Cipriano, Alexey Kazakov, Dmitry Ravcheev, Adam Arkin, Mikhail Gelfand*, Inna Dubchak* (gelfand@iitp.ru, ildubchak@lbl.gov)
Genomics Division, Lawrence Berkeley National Laboratory, The Virtual Institute of Microbial Stress and Survival, The Research and Training Center on Bioinformatics (Moscow)

ESPP2 Environmental Stress Pathway Project

DOE GENOMICS:GTL ACCELERATING DISCOVERY FOR ENERGY AND ENVIRONMENT

VIMSS Virtual Institute for Microbial Stress and Survival

http://vimss.lbl.gov/

INTRODUCTION

http://regtransbase.lbl.gov

RegTransBase, a database describing regulatory interactions in prokaryotes, has been developed as a component of the MicrobesOnline/RegTransBase framework successfully used for interpretation of microbial stress response and metal reduction pathways. It is manually curated and based on published scientific literature. RegTransBase describes a **large number of regulatory interactions** and contains **experimental data** which investigates regulation with known elements. Currently, the database content is derived from more than 4500 relevant articles describing over 10000 experiments in relation to 180 microbes. It contains data on the regulation of ~17000 genes and evidence for ~8800 interactions with ~970 regulators.

RegTransBase additionally provides an expertly **curated library of alignments** of known transcription factor binding sites covering a wide range of bacterial species. Each alignment contains information as to the transcription factor which binds the DNA sequence, the exact location of the binding site on a published genome, and links to published articles. RegTransBase builds upon these alignments by containing a set of computational modules for **the comparative analysis of regulons** among related organisms. These modules guide a user through the appropriate steps of transferring known or high confidence regulatory binding site results to other microbial organisms, allowing them to study many organisms at one time, while warning of analysis possibly producing low confidence results, and providing them with sound default parameters.

There is an increasingly tight coupling of RegTransBase with MicrobesOnline in reporting cis-regulatory sites and regulatory interactions, and integrating RegTransBase searches into MicrobesOnline cart functions.

RESULTS

Data SUMMARY

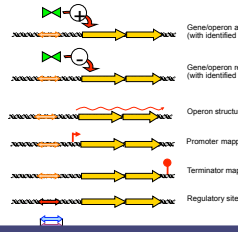
Experiments - taxonomy distribution

Bacteria	9347	
Proteobacteria	6615	
Alphaproteobacteria		2385
Betaproteobacteria	315	
Gammaproteobacteria	3843	
Delta/Epsilon subdivision	199	
Firmicutes		2003
Bacilli		1817
Clostridia	187	
Cyanobacteria	486	
Actinobacteria	269	
Archaea	345	
Plasmids/Transposons	538	
Phages	162	

Number of experiments

Single genome experiments	9236
Multiple genome experiments	978

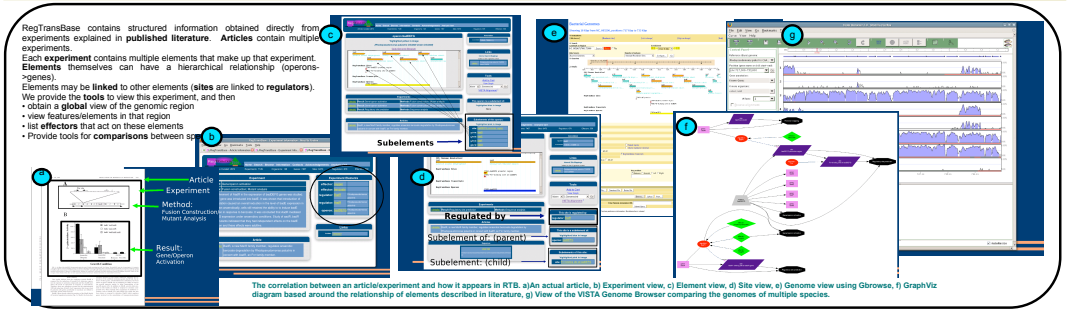
A decision of whether to include each putative site in a particular regulon is made after consultation with scientific literature by a human expert. RegTransBase (RTB), a manually curated database of regulatory interactions, captures the knowledge in literature using a controlled vocabulary. RTB contains the following types of experimental data:



MATERIALS

Information

RegTransBase contains structured information obtained directly from experiments explained in **published literature**. **Articles** contain multiple experiments.
Each **experiment** contains multiple elements that make up that experiment. **Elements** themselves can have a hierarchical relationship (operons->genes).
Elements may be **linked** to other elements (**sites** are linked to **regulators**).
We provide the **tools** to view this experiment, and then
- obtain a **global** view of the genomic region
- view features/elements in that region
- list **effectors** that act on these elements
- Provide tools for **comparisons** between sp



The correlation between an article/experiment and how it appears in RTB. a)An actual article, b) Experiment view, c) Element view, d) Site view, e) Genome view using Gbrowse, f) GraphViz diagram based around the relationship of elements described in literature, g) View of the VISTA Genome Browser comparing the genomes of multiple species.

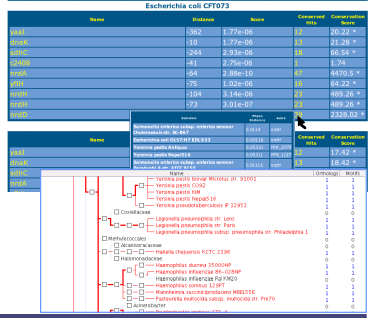
RESULTS

Comparison

RTB has a manually curated collection of over 100 position weight matrices and alignments (with plans for more in the future). We provide the ability to search sequenced genomes using these matrices or the user can supply their own alignment. Using a collection of interfaces we aim to provide a tool for the following situations:

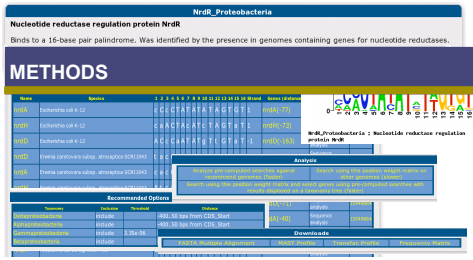
- One matrix + one genome of interest
  - Show predicted binding sites which match this matrix, while providing additional information.
- One gene + multiple genomes
  - Predict binding sites for orthologous genes using certified matrices.
- One matrix + multiple genomes
  - Compare the predicted binding sites across genomes for a particular matrix, highlighting orthologous similarities.
- Multiple matrices + multiple genomes
  - Compare the predicted binding sites across genomes for a set of matrices.

These tools allow a person to be guided through a semi-automated process which will highlight conserved transcription factor binding sites.



CONCLUSIONS

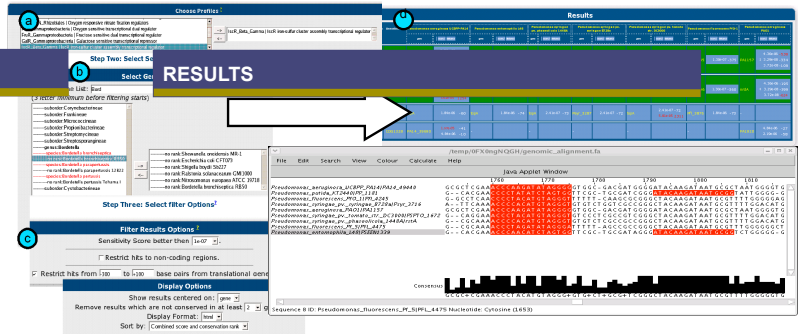
Alignments of Binding Sites



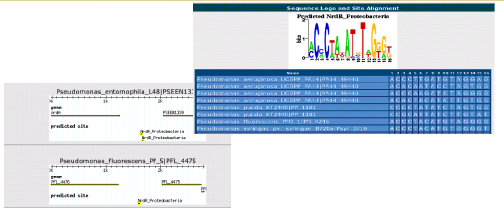
METHODS

In addition to publication data, RTB provides its users with a growing collection of curated binding site alignments. Each alignment was created by an expert curator who provided descriptions explaining all alignments, specific sequence locations referenced to NCBI RefSeq genomes, available publications, and recommended options for using this alignment to search new genomes. This data is available for download

Prediction



RESULTS



The process for comparing hits of a particular motif against multiple genome. a) A predefined alignment is chosen to create a position weight matrix from (custom alignment option is also available), b) Genomes to compare are selected, c) Results will be filtered by the options given, d) The result is a table with rows being orthologous genes, and hits specified within each row. For each orthologous row, additional analysis tools are available, such as sequence logos, sequence alignments in graphical and text formats, phylogenetic trees and the ability to view the alignment in the feature rich application JalView.

ACKNOWLEDGEMENT

ESPP2 (MDCASE) is part of the Virtual Institute for Microbial Stress and Survival (VIMSS) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics:GTL Program through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.